NASA CONTRACTOR REPORT

NASA CR-257

NASA CR-257



# MEASUREMENT CRITERIA IN MAN MACHINE SYSTEMS SIMULATION

*by R. W. Obermayer*

Prepared under Contract No. NASw-869 *by*
BUNKER-RAMO CORPORATION
Canoga Park, Calif.
*for*

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION • WASHINGTON, D. C. • JULY 1965

NASA CR-257

# MEASUREMENT CRITERIA IN MAN MACHINE SYSTEMS SIMULATION

By R. W. Obermayer



Prepared under Contract No. NASw-869 by
BUNKER-RAMO CORPORATION
Canoga Park, Calif.

for

NATIONAL AERONAUTICS AND SPACE ADMINISTRATION

For sale by the Clearinghouse for Federal Scientific and Technical Information
Springfield, Virginia 22151 – Price $1.00

TABLE OF CONTENTS

# MEASUREMENT CRITERIA IN

# MAN MACHINE SYSTEMS SIMULATION

By R. W. Obermayer

27397

## SUMMARY

This report describes simulation, models and games as analogies. They resemble in some way something else about which information is desired. We may therefore measure an analogy instead of the real-world object. Critical dimensions of analogies are the level of abstraction and the fidelity of simulation, however, if the object is to measure, the most critical aspect is the validity of measurement. Unfortunately, validity is not always a practical concept. Since the objective of measurement is to derive information, simulation studies are analyzed with respect to information objectives in the attempt to derive criteria for measure selection.

## INTRODUCTION

The ability to measure is frequently used as a gauge of the maturity of scientific disciplines. This is a reasonable gauge since upon reflection it will be seen that what we understand through research depends upon measurement, and what we can predict in the design of systems also depends upon what we have measured. Consequently, it is virtually impossible to overestimate the importance of measurement.

Examination will show that much of man-machine measurement depends on techniques variously known as simulation, modeling or gaming. With the hope of contributing to the cause of better measurement, in the following, the theoretical foundations of simulation, models and games are outlined. With the view that measurement is for the purpose of gaining information, the applications of these techniques are examined with respect to information categories. Some of the problems concerning this discussion are simulator selection, measure selection and measure validity.

# THEORY

## Definition

A variety of definitions of games, models and simulation exist in the man-machine systems literature. Abt (1964) distinguished between games, models and simulations as follows: "A game is any contest played according to rules and decided by skill, strength, or apparent luck. A model is a representative -- actual or theoretical -- of the structure or dynamics of a thing or process. A simulation is an operating imitation of a real process." It is immediately recognized that he does not consider these as mutually exclusive categories. While no small amount of difficulty would be met in clarifying and extending these definitions, the problem is compounded since there is little consistency in the literature. Flagle (1960) defines: "By simulation is meant the technique of setting up a stochastic model of a real situation, and then performing sampling experiments upon the model." "With this definition, simulation may be regarded as one of several forms of applications of Monte Carlo techniques." With regard to system engineering, Goode and Machol (1957) state: "...we shall here define simulation to be the study of a system by the cut-and-dry examination of its mathematical representation by means of a large-scale computer." Without proceeding further it is clear that various disciplines using similar techniques apply somewhat different terminology.

One other attempt at definition will be helpful. Chapanis (1961) proposes the following: "Models are analogies. Scientific or engineering models are representations, or likenesses, of certain aspects of complex events, structures, or systems, made by using symbols or objects which in some way resemble the thing being modeled." This definition goes to the heart of the matter and will be basically the definition used here. Here we shall say that games, models and simulations are analogies, and shall not attempt to distinguish between them. Abt's definition is indicative of the differences in the approaches that one might adopt, but more may be gained by pointing up the basic sameness than by perseverating on the assignment of labels. In view of the existing confusion of terminology, the terms, game, model or simulation, will be used interchangeably but with some attempt to be consistent with a specific literature.

There are a wide variety of reasons for manufacturing analogies of the real world, but one of the most important is to create an environment which permits measurement. There are of course important uses of simulators not requiring formal measurement, such as the training of complex skills, but generally even with these there exists at least a secondary requirement for measurement (e.g. proficiency measurement). This paper will only consider simulations, models and games construed for the purpose of producing information through some level of measurement -- in short, any scientific analogy which permits measurement.

## Levels of Abstraction

In constructing an analogue of a real world situation one may choose from a spectrum of symbolic representations. One description of this spectrum is given by Haythorn (1962), shown in Figure 1. The real world is depicted at one end of the spectrum and the mathematical model at the other end. In between are analogues of different levels of abstraction. As the corresponding models generally develop, as one proceeds from bottom to top in the figure, the models increase in abstraction, symbolization and generality; and commonly the models decrease in validity and in the amount of real-world detail represented. However, it is important to note that the degree or level of abstraction is a unique property of the analogy and is separate from such matters as the amount of real-world detail (which will be called "Fidelity of Simulation" below). It may also be noted that techniques corresponding to the level of abstraction at the bottom of the figure involve collection of basic data, while at the top the techniques involve synthesis of basic data. Therefore, the choice of analog depends greatly on available knowledge of the phenomena to be studied. If the available knowledge is scanty, the analog should bear the closest possible resemblance to the real world; if available knowledge of the phenomena is complete, measurement merges into calculation through a mathematical model.[1]

A fundamental problem is that the real world is only displayed through observation and measurement. As a result, the analogies which we construct are based on these imperfect descriptions. Further, a common reason for measurement based on an analogy is that the real world situation in question does not yet exist. Consequently, generalizations based on analogy must always be suspected.

## Fidelity of Simulation

At any given level of abstraction an analogy of the real world may represent only a selected subset of the real world detail, and that detail may be included with varying degrees of precision. For example, a flight simulator may include only the static response characteristics (airspeed, altitude, etc.) or it may include the full dynamic characteristics (oscillatory transients occurring in the change from one state to another). The characteristic of comprehensiveness and precision of simulation is referred to as the fidelity of simulation.

---

[1]It may be questioned whether a computer output, on a completely programmed basis, may constitute measurement. It does if one applies the definition of Stevens (1951): "In its broadest sense measurement is the assignment of numerals to objects or events according to rules."

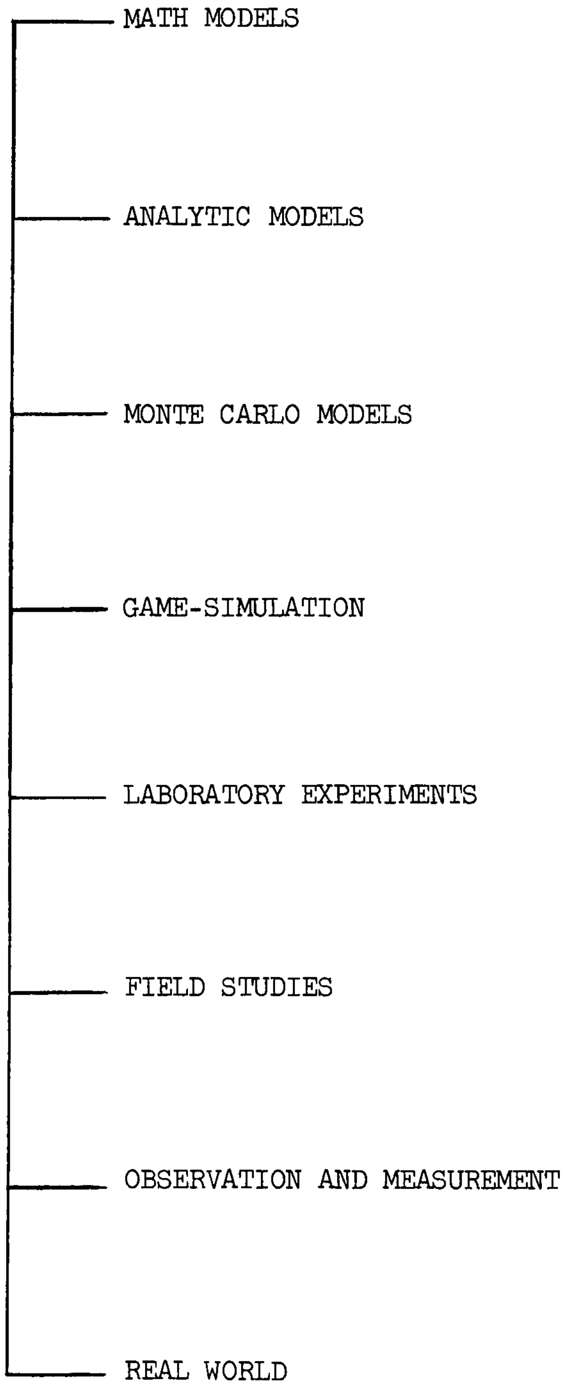


FIGURE 1. Levels of Abstraction

The question of fidelity of simulation also poses a dilemma. One is ordinarily more confident of his results with high fidelity of simulation, but frequently simulation is of greatest value when real-world aspects are missing, forcing low-fidelity of simulation. For example, many problems are studied with simulators because hazardous real-world attributes are deleted. Additionally, high fidelity simulation may be extremely expensive.

When the object of the simulation is to measure, one approach to this dilemma is to use the minimum of real-world detail to produce valid measurement. If the measurements on the analogue correlate with real-world measurements, high fidelity is an unnecessary, perhaps undesirable, complexity. The corresponding problem of fidelity of flight simulation for training has been extensively analyzed by Muckler, et al. (1959). The criterion in this case is the positive transfer of training to the ultimate flight vehicle. Muckler gives evidence that sometimes simple trainers of procedures are adequate, whereas at other times, aspects of the flight dynamics which may be judged as "small" are required to cause positive transfer of training. Indeed, the increase of fidelity by additional flight dynamics, but with a poor approximation, may cause negative transfer. In a study by Brown, et al. (1958) of the centrifuge as a flight simulator it was concluded: "For the simple tracking tasks employed in the present experiment, the results of work with a static, or fixed-base, simulator provided just as good a basis for prediction of the way in which pilots would perform a specific task in the aircraft as did work performed on the centrifuge." In fact there was some suspicion that anomalous rotations of the gondola might have elicited negative effects.

The problem is more complex than the difficult one of deciding on the proper amount of real-world detail to be incorporated in the analogy to produce valid measurement. It may be necessary to include deliberate distortions into the model to derive valid measurement. This may be demonstrated through an example: The areas and volume of an aircraft model depend upon the square and the cube of its dimensions, respectively. Since the aerodynamic properties of the model depend on area and volume, a scale model -- one that looks just like the real aircraft -- will produce erroneous wind-tunnel data, while a distorted-appearing model will produce valid measurement. It may be seen that one must be careful in which aspects he requires his analogies to "look" like real world objects.

The question of fidelity of simulation is complex, requiring separate study in each specific case. If one must simulate as a short-cut in lieu of an understanding of the phenomena simulated, it will be difficult to answer basic questions of fidelity of simulation.

## Measurement Validity

When one simulates to gain information through measurement, the success of the endeavor ultimately depends on measure validity. It will be seen however that this is a somewhat circular statement. Crudely put, valid measures mean what they are supposed to; in this vein, valid measurement is synonymous with successful simulation.

To be more precise the following definitions of measure validity were recast from those existing in the literature (cf. McCoy, 1963; Smode et al., 1962; APA Committee on Test Standard, 1954).

1. Predictive validity. The degree to which a measure derived from simulation correlates with the same measurement taken in the real-world environment.

2. Concurrent validity. The degree of agreement between two different measures simultaneously taken in the same environment.

3. Content validity. The degree to which a measurement taken in the simulated environment incorporates all factors necessary to predict conditions of the real world.

4. Construct validity. The degree of correlation between a given measure and some construct, i.e. that certain explanatory constructs account for measurement values.

These definitions can perhaps be made clearer through reference to Figure 2. Validity is normally measured in terms of the correlation between repeated measurements of two measures. In the figure, the degree of predictive validity involves a comparison of measure A (from the analogy) against measure C (from the real world). Concurrent validity is determined by a comparison of measure A and measure B, both derived from measurement on the analogy. Content validity is indicative of comprehensiveness of measurement, and is measured by the ability to predict some real world condition or event. Content validity is therefore commonly based on a battery of measures, and may be judged by the ability to predict a condition such as mission success which may depend on measures of response, reliability, acceptability, etc. (measures A, A', A'',...compared to measure E). Construct validity involves the agreement between measurement and concept. In the figure, measure A may be expected to be an indication of operator workload, however, there is in general no measure D. In the sense that the validity is measured in terms of the correlation of two columns of numbers, it may be seen that only predictive and concurrent validities can be quantified, and that content and construct validity are largely subjective terms.

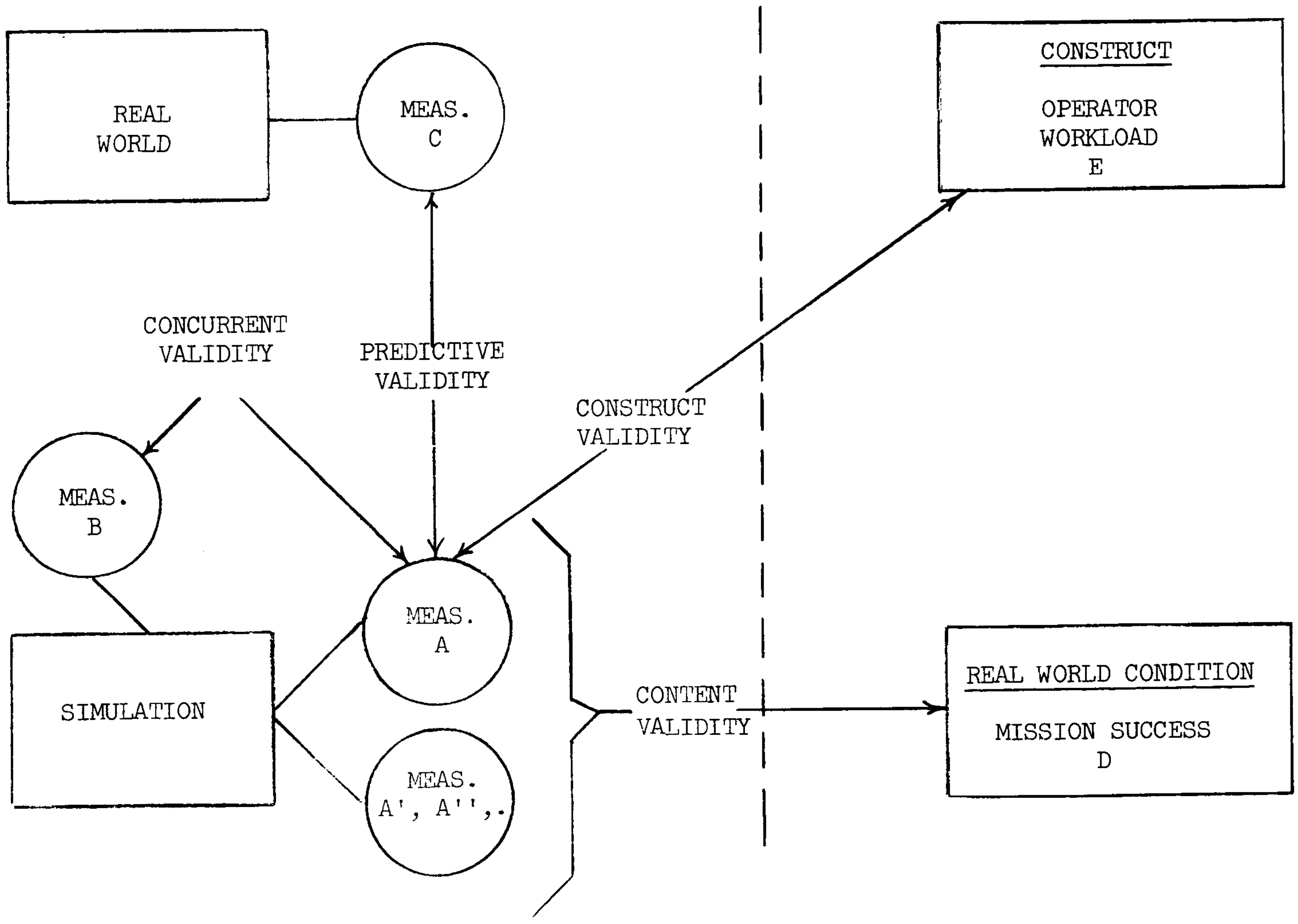


FIGURE 2. Four Types of Validity

Ordinarily, predictive validity is sought in simulation, with the goodness of simulation judged by the degree of correspondence of measures taken in the simulated and real worlds. Concurrent validity is of interest when substituting one measure for another. Content validity is of importance for assuring that measures predict abstract criteria. Construct validity is sought for validating theories, which may be valid independently of any agreement between simulated and real worlds.

It has already been pointed out that predictive validity is most commonly sought for validating simulation, however, this is not necessarily the best choice. Due to greater control, measurement in the simulated environment may be more reliable and therefore not well correlated with unreliable real-world measurement. Content validity may be frequently the only important factor; if goal achievement can be predicted, it does not matter if the measure correlates in the real world or is even measureable in the real world. A theory demonstrating high construct validity can be of great importance (although perhaps less useful) if it lacks relating measures of high predictive validity. Even where predictive validity is sought, for example, when comparing two aircraft subsystems in the flight simulator, one may be content with valid ranking of measures, rather than perfect agreement between simulated and real worlds.

Unfortunately simulation, models and games are not often validated. In a great many cases there is no alternative to the use of simulators, models or games. They afford the only available approach to certain kinds of problems and are used even though their real value is not accurately known. The danger involved is obvious; without validation, measurement may be the collection of worthless numbers.

Further, there is the danger of overgeneralizing from models. The models are analogies and cannot be completely accurate. Models, as opposed to theories, do not attempt to describe the thing that they represent. As Chapanis (1961) puts it, "Models, in a word, are judged by criteria of usefulness; theories, by criteria of truthfulness." Similar to the logical fallacy of regarding the premise to be true if the conclusion is true, the predictive validity of a model does not imply that the real world is like the model.

## APPLICATIONS

In brief, the previous discussion expands on the theme that simulators, models and games should be designed to generate valid measurement. But what measurement? The selection of measures is a difficult topic in itself, but the success of simulation, modeling and gaming is directly dependent upon it. In the attempt to explicate the problems of measure selection, in the following, a number of exemplary studies are analyzed with respect to measures selected and information derived.

## Simulation for Information

Clearly the objective of simulation as a measurement tool is to collect information. It may be reasonable therefore to attempt to distinguish between various simulation studies in terms of the particular information objectives. In the following, a number of simulation studies are cited and discussed. The majority of these studies were performed in the support of system design, development and test. For these studies the types of information sought at various times during the system development cycle serves as a framework for classification. Apart from the requirements of the system development cycle, simulation serves well as a method of research, and two research studies are discussed without further classification.

## Simulation in Support of System Design

Six basic types of simulation studies may be identified in terms of the information provided during the system design cycle: (1) early in system design simulation techniques may provide initial feasibility demonstrations, (2) in the early design stage, system models serve as an analytic technique, (3) simulation allows the detailed comparison of specific subsystems, (4) simulation provides a method for the collection of system design data and user criticism, acceptance and design advice, (5) simulation allows system test, and sometimes (6) total system performance evaluation. Each type of simulation study incorporates different methodology and different measurement requirements.

Feasibility demonstrations. Pilot participation in the control of large space boosters has been considered a potentially important role, but due to the stringent requirements of the task, a role requiring proof of its feasibility. To explore the feasibility of pilot control during boost, Muckler, Hookway and Burke (1962) studied the insertion of pilot control into a simulated booster in (a) the booster flight control loop, with and without the benefits of autopilot rate damping, and (b) the guidance loop, where the pilot attempted steering control by applying torque to the attitude gyros. The study was composed of two principal parts: (a) an analysis of total flight control loop stability using a mathematical model to approximate the pilot, and (b) an empirical evaluation in which several pilots flew the simulated booster at various flight conditions. The test method was to adjust the simulation to represent a specific point on the boost trajectory and to apply a disturbance.

The purpose of the study was to establish initial boundaries for pilot booster control, to exclude the most obviously unsatisfactory conditions from further consideration, and to recommend the most promising control modes for most stringent tests. The measurement emphasis, therefore, was on measures of stable performance and maintenance of safe vehicle tolerances. Measurement included, for example, maximum vehicle body rates, vehicle attitude error, measures for checking the mathematical model, and pilot performance on secondary tasks.

Based on these measurements, a number of conclusions were reached about the adequacy of pilot control. Stable pilot performance was probably inadequate without stability augmentation, and performance was a function of pilot loading and the pilot's position in the control loop. The mathematical model, although quite simple, was found to predict stability or instability.

It may seem that this study provided excellent preliminary information for the design of large space boosters, although clearly further design would require much more detailed analysis. It may be noted that this study wisely checked by direct empirical test any predictions of the simple mathematical model; conceivably, after sufficient empirical tests, the mathematical model of the human pilot may serve to expedite such feasibility tests.

Analytic system models. One of the more prominent signs of the impact of the modern high-speed digital computer on man-machine methodology is a technique called computer simulation. Computer simulation consists of constructing a mathematical model of each system element (which need not be a concise mathematical statement, but may consist of tables of values, if... then statement, probability distributions, etc.), and including all known interactions. The computer is then programmed to perform the indicated mathematical activities. The computer can repeat the activity many times under varying circumstances, observe itself, and produce a printed summary. Through this technique the complex system interactions can be observed while salient system properties are changed. The technique is of course directly dependent upon the accuracy of the mathematical model, however, it is mainly the model form which is critical (uncertain parameters are less of a problem since they may be varied to observe system sensitivity to a range of values).

Siegel and Wolf (1963) describe a technique designed to determine whether a two-man team can be expected to complete all actions required for a given time-dependent task within time limits. Through this technique the designer can determine task or system probability of success, operator loading, the distribution of failures, and the efficiency of the team workload division. Of course the model demands highly specific subtask data, such as: average subtask execution time, and the corresponding distribution, subtask probability of success and priority, sequencing of operation with necessary waiting and idling data, requirements for communication, etc. In brief, highly specific task activity information is required. Fortunately, these data largely exist or may be readily measured.

The basic logic of this technique has been adapted to four tasks: carrier landing, inflight missile launching, in-flight refueling, and in-flight intercept. With regard to the success which one might expect, Siegel and Wolf comment: "For all four tasks reasonable concordance was found between the predictions from the model and outside criteria of success on the task involved." "At present it appears that for systems similar to those tested, reasonable predictive efficiency may be anticipated for the model."

In the utilization of analytical stochastic models to investigate alternative system configurations, it is possible -- indeed critically necessary -- to adapt a thorough experimental design (cf., Ruby et al., 1963). In a study of alternative ways of organizing a logistics support system, Haythorn (1962) used a design controlling for: 2 management structures, 2 weapons systems, 4 stress conditions, 64 parts and 9 bases. The total design was a complete factorial with the latter three variables further defined by Greco-Latin arrangements. Operation of the system occurred within the computer managed by Air Force logistics experts who participated as subjects in the experiment. The Air Force personnel received information regarding the performance of the system, made management decisions, and implemented the decisions by communicating to the computer model. Data collection and analysis were programmed on the computer allowing completion of an analysis of variance every simulated week of the study. The primary measure of performance was the occurrence of stock-outs, or demands for spare parts not available at the base.

The computer simulation technique normally includes simulation of probabalistic factors. Therefore whether real or simulated human elements are present to contribute to chance occurrences, experimental control is required to assure statistically significant results. As Haythorn remarks, "...even in systems as complex as Air Force logistics systems, it is possible to construct experimental designs that control stimulus variables and that such designs increase the predictability of one's results."

Subsystem comparisons. A problem which frequently occurs in system development involves the need to make choices among available hardware items based on complex decision-making criteria. This is perhaps the most common simulator application and consequently incorporates the most highly developed methodology.

The principal problem in using the simulator for subsystem comparisons is to assure test over the full useful range of the hardware with valid procedures, careful experimental control and adequate measurement of both subsystem and system performance.

An example of the methodology is provided by Gainer and Brown (1961). Highly experienced test pilots flew a flight simulator over a standardized mission profile using, in succession, three different altimeters (with statistical control for order of presentation). The goal of the study was to evaluate the three altimeters. The following maneuvers were flown in continuous sequence: (a) take-off and climb to 40,000 feet, (b) 180° turn to left, (c) 180° turn to right, (d) straight and level hold, (e) descent to 20,000 feet, (f) climb to 27,000 feet, (g) descent to 4,000 feet, (h) straight and level hold, (i) climb to 18,000 feet, (j) 180° turn to right, (k) 180° turn to left, (l) jet penetration and (m) low approach. During each maneuver, one-minute scoring periods were taken for heading, altitude, mach, vertical rate, and airspeed, where the particular measure was appropriate. Three kinds of measures were obtained: system performance measures, pilot preference measures, and reading errors, i.e., indications of expected system performance, user acceptance, and subsystem performance.

In contrast to the requirements for feasibility demonstrations and parametric analyses, specific hardware comparisons are usually conducted to make a firm and final decision. Of course the final decision must depend, in addition to the simulator data, on considerations of weight, space, cost, reliability, maintainability, safety, etc.

It may also be apparent that the results may differ from the study cited if different procedures were followed. Much depends on the manner which the simulation hardware is used, and thus no flight simulator per se can be validated; measurement validity depends upon the entire methodology.

Design advice and user acceptance. Many systems are designed to be used by people who are intelligent, skilled, experienced, and in positions of authority and with the responsibility of accepting or rejecting the final system -- and in some cases, partially capable of designing the system. During system design and development, when a simulator of sufficiently low abstraction exists, the system designers may wish to collect specific comments and opinions from the prospective users based on simulated experience. In this way he may collect valuable design advice based on the most appropriate subject population, and he may avoid some problems due to user rejection which may not otherwise be apparent prior to final acceptance testing.

User opinion data has been a tradition in aircraft development; Belsley (1963) considers this type measurement to be critical for all levels of flight simulator testing. In part, these data consist of unprompted opinions and questionnaire replies; in part these data consist of concerted attempts to correlate pilot opinions with aerodynamic parameters. A body of data called handling qualities requirements consists of systematic collection of pilot opinion (quantified on a ten-point scale) as a function of period and damping characteristics for longitudinal (short period and phugoid) and lateral dynamics. One can then plot opinion contours on a graph of period vs damping, and use this information to determine if a given design will meet with pilot acceptance, i.e., have satisfactory handling qualities.

Whereas most measurement would be appropriate for simulation at any point of the spectrum of abstraction, it is clear that information from the user can only be collected for low-abstraction simulation, i.e., close to the real world in regard to the user's task.

System test. Final system performance evaluation should naturally be conducted with the full real system. However, in a great many cases the tests performed with the simulated system constitute, for all practical purposes, the final system test. In some cases the system cannot be tested (e.g., may require all-out war), and in other cases the cost of system failure is so great that it simply cannot be permitted to remain undiscovered until a testable system is available.

Some excellent examples are provided by Chambers (1963) with regard to the Aviation Medical Acceleration Laboratory (AMAL) centrifuge tests of the X-15, Mercury, Dyna-Soar and Apollo vehicles (and for examples and discussion of ground systems evaluation, see Davis and Behan, 1962). The primary characteristics of system test are high-fidelity, low-abstraction simulation with profuse measurement. At AMAL, for example, the measurement equipment consisted of multichannel recorders, magnetic tape recorders, closed loop TV, and extensive analog and digital computer data reduction equipment. With the system being freely exercised in system test, experimental control is at a minimum, and one must be prepared to record and analyze achievement of goals and subgoals, system and all subsystem behavior -- virtually any contingency. The problems then are not usually those of defining the mode of simulation -- since the simulation usually approximates the real world as closely as possible -- but of providing sufficiently a broad spectrum of measurement.

## Simulation for Research

Use of simulation is of course not limited to support of system development. As a matter of fact it is probably more appropriate to system research efforts due to the inherent possibility for control and measurement. System research is extensively rich in complexity, consequently no attempt will be made to illustrate its nature fully. Only two examples will be given: both are research of human behavior concerning widely different topics.

Models for human tracking behavior. As several of the previous examples document, the flight simulator provides a control task for the human operator, with the simulator providing stimuli to the operator and responding in a closed-loop fashion to the operator's responses. The continuous tracking behavior of the human operator is complex and has so far defied description in comprehensive terms.

While some models exist which will approximate the gross aspects of human operator response, a study by Adams and Webber (1963) attempts to derive a model whose output will match specific features of the operator's output. Using analog computer techniques to simulate a generalized control system, data were collected upon which a digital-computer-produced model was based.

The model attempts to reproduce stochastic features of human tracking behavior utilizing the notion of an error peak; tracking error rises tc a point where it is sensed by the human operator and then is reduced by the operator to his criterion of excellence; further system disturbances cause this cycle to repeat. The data of four groups of 12 subjects were used to compute distributions related to tracking error peaks and to predict tracking time histories for four groups of 12 hypothetical subjects. Encouraging agreement was found between the data of the hypothetical and real subjects.

The measures recorded consisted of data for model computation and measures of model success. The specific measures were: sampled digital tracking errors, time-on-target scores, frequency distributions of error peak values, mean number of peaks, and frequency distribution of time intervals between peaks. In passing, it may be noted that this particular set of measures is probably typical of no other simulation study.

Business gaming. An example of gaming which provides a foundation for research is given by Kennedy's (1962) account of Project SOBIG at Princeton University. In a business game model based upon the stock exchange, 42 three-man teams were studied under various conditions of inter-team competition and intra-team cooperation. The teams acted as the investment committee of banks. Within the environment of a four-stock-and-bond "supply-and-demand" market, 10-13 years of stock market operation were played out in a period of three months. The objective of each team was to make as much money as possible. The team members were provided with the task of processing a continuous stream of information as well as planning requirements for long-range decisions. In addition to the subject teams, several model teams (unmanned) with fixed programs were inserted into the competition against the manned teams.

In general, the measurements consisted of the transactions and accumulated gains of each team, with comparisons made across teams and in particular against the model teams with fixed, controlled strategies. The results gave some new insights into long-term planning behavior. The surprising finding was that the long-term planning employed by the subject-teams was generally inferior to the strategy of making no transactions at all.

## DISCUSSION

### Selection of Simulation Device

With just the sampling of the literature viewed in the previous sections, it should be apparent that there is an extreme variety of simulation devices used in practice. In the area of flight simulation alone, devices range from simple photo mockups and control system mockups to full dynamic simulation with six degrees of freedom. It is interesting to speculate as to a choice of simulation if one is faced with the problem of measuring pilot-aerospacecraft performance. The range of possible choices are illustrated in Figure 3. One may seek a choice of a simulator along two dimensions: The relations observed in the real world environment may be abstracted into other symbolic terms, and, the relations observed in the real world may be simplified or ignored, i.e., decrease fidelity. For example, fidelity may be nearly maintained, but with a large change in the degree of abstractness, by representing the full aerodynamic complexity in the form of extensive and complicated equations. The degree of abstractness may be maintained at a relatively low level, but with a great reduction in fidelity, by using a small, light plane. Other possible simulator choices range from simple photo-mockups to highly complex centrifuge-computer-cockpit-mockups.

There are ordinarily a number of practical considerations in choosing a simulator. The investigator normally wants a measurement environment he can control, convenience and economy.

Thus there are many practical considerations which, of course, cannot be ignored. As a matter of fact it would appear that these considerations have had a dominating influence on simulator selection. However, the decision should not be made solely on these bases, since the main item is the measurement objective of the simulation.

Here we are assuming that one simulates in order to measure. The first consideration should be therefore the validity of the measures derived through simulation. Collection of invalid data is a waste of time.

It may be seen that there is no direct relation between validity and fidelity/abstraction. Whether the aircraft is measured or highly complex equations are measured, valid data may result. Tests of the pilot's procedural skills may be just as valid in the low-fidelity photo-mockup

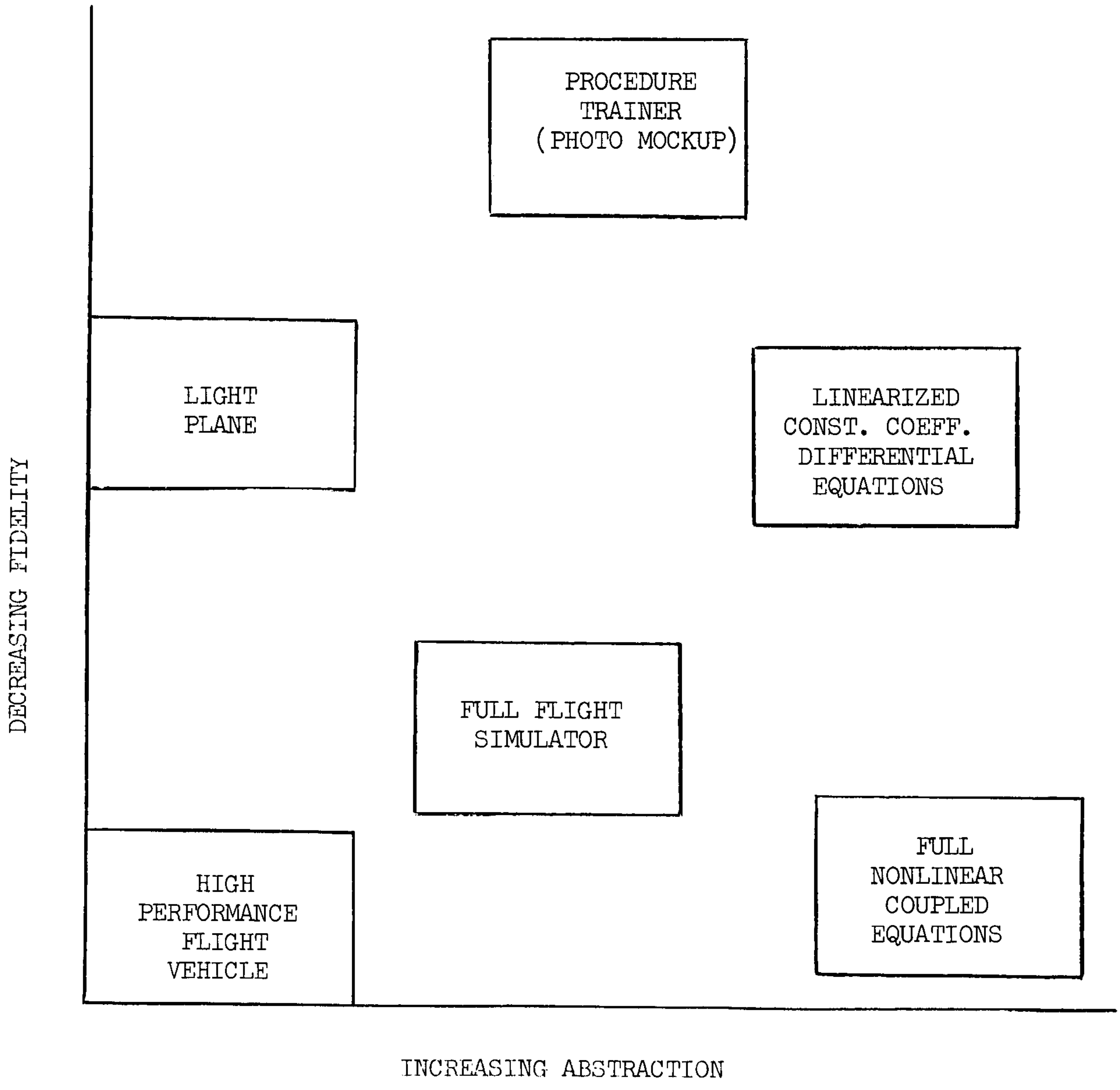


FIGURE 3. Simulation as a Function of Abstraction and Fidelity

as in the high performance aircraft. For that matter, the question is relative, for if one wishes to predict light plane performance, the high-performance aircraft would be considered a low-fidelity representation and may yield invalid measurement.

Therefore, no specific guide can be given for simulator selection except to underline the requirement for valid measurement. Parenthetically, it may be mentioned that even the usual requirement for experimental control may be a disadvantage. While one measures in the field because of simulation complexity and cost, a distinct advantage of field test is the potential for measurement of events which would be suppressed by the usual experimental controls.

## The Need for Validation

As defined here, simulation, models and games are analogies. They resemble in some way something else about which information is desired. We must remember when we measure the analogy that the resemblence between analogy and the real world object can never be complete. The potential for erroneous measurement always exists unless the validity of the analogy has been demonstrated. Unfortunately few of our models have been validated, nor is the situation likely to improve since we usually do not have the time.

Moreover, the problem is not alleviated by noble resolutions since validity is not always a very practical concept. Frequently the real world system being modeled does not yet exist, precluding any predictive validity checks. Frequently our ability to measure is not as good in the real world environment as it is in our simulations, so how do we validate measures? The validity we seek may be with respect to the success of achieving goals which are measureable only in retrospect. To continue, to validate may require more trouble than measuring in the real world; to get the job done within time limits requires taking some chances.

On the other hand, when one does not validate, he must have a great deal of faith in how own powers of analytic and synthetic reasoning.

## Measure Selection

It is strikingly apparent in scanning the simulation literature that a wide variety of measures are in use. The scope of information requirements under the rubrics of simulation, models and games certainly is very broad, and the situation is not much better when the topic is narrowed to just flight simulators (cf., Obermayer and Muckler, 1963).

Some general statements may be made with regard to the information category one wishes to measure (see Table 1). To demonstrate feasibility, it is necessary to show that some simple tasks can be accomplished and that no unsafe conditions result. In analytic efforts, one may vary system parameters (conduct an experiment) with the primary measure being system performance. Moreover, it is also necessary to measure the immediate efforts on the subsystems directly affected and the potentially adverse effects on other subsystems. Similarly, with subsystems comparisons, system and subsystem performance, and frequently user acceptance are measured. System test depends primarily on the measurement of acceptable system performance through objective and user acceptance measurement, but one also must provide for malfunction data to facilitate design improvements. If one is interested in model building, there is a need to provide data for computing or double-checking model parameters and the comparison data to the real world being represented. In general, it is believed that if the measurement categories shown in Table 1 can be delimited, the specific measurement requirements will be very apparent.

It would appear, then, that when one can state what it is that he wants to know, specific measurement is implied. It might be expected therefore that some standardization would exist across similar studies. However, no such standardization is readily apparent, although a degree of standardization would certainly facilitate prediction and measure selection, and, in time, would alleviate some concern about validity of simulation.

TABLE 1

SELECTION OF MEASUREMENT

| INFORMATION CATEGORY | MEASURE |
|---|---|
| Feasibility | Stability, Safe Performance |
| Analysis | System Performance<br>Subsystem Performance<br>Acceptance |
| Subsystems comparison | System Performance<br>Subsystem Performance<br>Acceptance |
| System Test | System Performance<br>Subsystem Performance<br>Acceptance |
| Model Building | Real World Input-output<br>Model Input-output<br>Model parameters |

## Quantification of Human Performance

Part of the problem in design of measurement is that we do not understand the basic phenomena well enough to specify what we wish to know. On the other hand, with respect to human performance, this is at times precisely the reason for simulation. Without the advantage of observing the human operating in a simulated environment, insufficient human operator data inhibits prediction of system performance. One of the potential evils of simulation is that given some ability to predict system performance, the measurement of corresponding human behavior is forgotten. The next time the same problem arises, one must again simulate because of insufficient basic data. On the other hand, simulation at the same time offers the greatest potential for the quantification of human performance.

It should be clear from the examples of simulation, modeling and gaming given earlier, that these techniques are applicable at every point of the system development cycle. Further, these techniques are used at various levels of abstraction by virtually every discipline which contributes to system development. There is much to be gained through the merging and refinement of techniques of all disciplines concerned with system effectiveness.

## REFERENCES


Abt, C. C. War gaming. International Science and Technology, 1964, 32, 29-37.

Adams, J. A., and Webber, C. E. A Monte Carlo method of tracking behavior, Human Factors, 1963, 5, 81-102.

Belsley, S. E. Man-machine system simulation for flight vehicles, IEEE Trans. Human Factors in Electronics, 1963, 4, 4-14.

Brown, J. L., Kuehnel, H., Nicholson, F. and Futterweit, A. Validity of the centrifuge as a flight simulator, NADC Report No. 4, USN Aviation Medical Acceleration Laboratory, Johnsville, Pa., 1958.

Chambers, R. M. Operator performance in acceleration environments. In Burns, N. M., Chambers, R. M. and Hendler, E. (Eds.) Unusual environments and human behavior: Physiological and psychological problems of man in space, London: Free Press of Glencoe, 1963.

Chapanis, A. Men, machines, and models. Am. Psychologist, 1961, 16(3), 113-131.

Davis, R. H., and Behan, R. A. Evaluating system performance in simulated environments, In Gagne, R. M. (Ed.) Psychological principles in system development, New York: Holt, Rinehart and Winston, 1962.

Flagle, C. D. Simulation techniques. In Flagle, C. D., Huggins, W. H., and Roy, R. H. (Ed.) Operations Research and System Engineering, Baltimore: Johns Hopkins Press, 1960.

Gainer, C. A. and Brown, J. E. Simulator test of Kollsman drum-pointer altimeter, counter-drum pointer altimeter, and Specialties altimeter. Contract AF33(616)-7752, Martin ER 11,787, Martin Company, Baltimore, Maryland, 1961.

Goode, H. H. and Machol, R. E. System Engineering; New York: McGraw-Hill, 1957.

Haythorn, W. W. System simulation as a technique in systems research. In Bennett, E., Degan, J. and Spiegel, J. (Eds.), Human Factors in Technology, New York: McGraw-Hill, 1963.

Haythorn, W. W. Information systems simulation and modeling. Paper presented at First Congress on the Information System Sciences, Hot Springs, Virginia, November 19, 1962.

Kennedy, J. L. Environment simulation as a technique for studying human behavior. Paper presented at the First Congress on the Information System Sciences, Hot Springs, Va., November 19, 1962.

McCoy, W. K. Problems of validity of measures used in investigating man-machine systems. Human Factors, 1963, 5, 373-377.

Muckler, F. A. Nygaard, J. E., O'Kelly, L. I. and Williams, A. C. Jr. Psychological variables in fidelity of flight simulation. WADC TR 56-369, USAF Aero Medical Laboratory, Wright-Patterson Air Force Base, Ohio, 1959.

Obermayer, R. W. and Muckler, F. A. Performance measurement in flight simulation studies. Paper presented at the AIAA Simulation for Aerospace Flight Conference, Columbus, Ohio, 26-28 August 1963.

Ruby, W. J., Jocoy, E. H., and Pelton, F. M. Simulation for experimentation: a position paper. Paper presented at the AIAA Simulation for Aerospace Flight Conference, Columbus, Ohio, 26-28 August 1963.

Siegel, A. I. and Wolf, J. J. Computer simulation of man-machine systems. In Burns, N. M., Chambers, R. M., Hendler, E. (Eds.). Unusual environments and human behavior: Physiological and psychological problems of man in space. London: Free Press of Glencoe, 1963.

Smode, A. F., Gruber, A. and Ely, J. H. The measurement of advanced flight vehicle crew proficiency in synthetic ground environments. MRL-TDR-62-2, USAF 6570th Aerospace Medical Research Laboratories, Wright-Patterson AFB, Ohio, 1962.

Stevens, S. S. Mathematics, measurement and psychophysics. In S. S. Stevens (Ed.), Handbook of Experimental Psychology, New York: Wiley, 1951.